10/829,613 PTO-892

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G06F 17/30 — A1

(11) International Publication Number: **WO 99/31606**

(43) International Publication Date: 24 June 1999 (24.06.99)

(21) International Application Number: PCT/US98/26642

(22) International Filing Date: 15 December 1998 (15.12.98)

(30) Priority Data:

08/990,316	15 December 1997 (15.12.97)	US
09/038,808	11 March 1998 (11.03.98)	US

(71) Applicant *(for all designated States except US)*: MANNING & NAPIER INFORMATION SERVICES [US/US]; 1100 Chase Square, Rochester, NY 14604 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: STOFFEL, Kilian [CH/CH]; Rue des Verenes, CH–2013 Columbier (CH). WOOD, Robert, L. [US/US]; 940 Five–Points Road, Rush, NY 14543 (US).

(74) Agents: KULAS, Charles, J. et al.; Townsend and Townsend and Crew LLP, 8th floor, Two Embarcadero Center, San Francisco, CA 94111–3834 (US).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

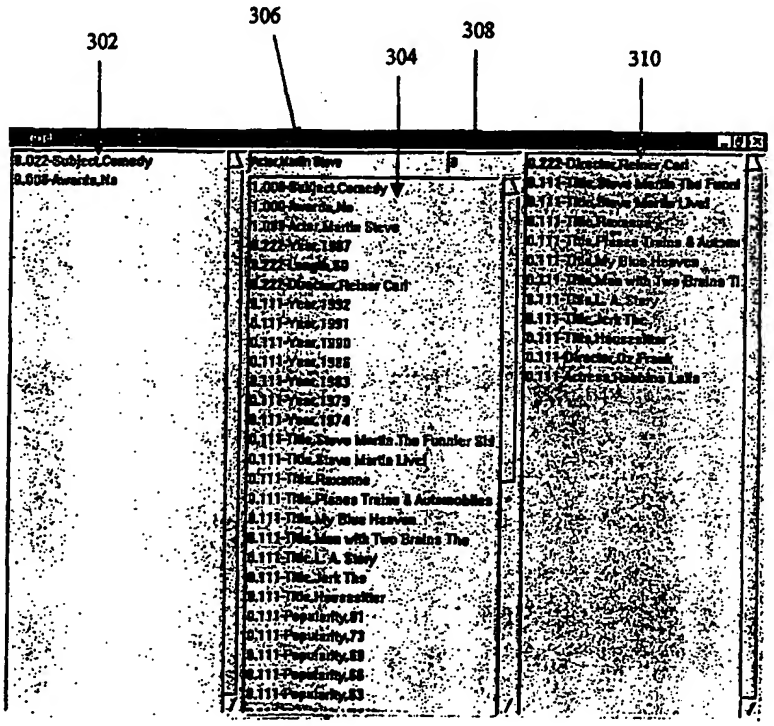
Published

With international search report.

(54) Title: DATABASE ANALYSIS USING A PROBABILISTIC ONTOLOGY

(57) Abstract

A method and system for efficiently analyzing databases. In one embodiment, the invention is used to analyze data represented in the form of attribute–value (a–v) pairs. A primary step in building the ontology is to identify parent, child and related a–v pairs of each given a–v pair (306) in the database. A parent (302) is an a–v pair that is always present whenever a given a–v pair is present. A child is an a–v pair that is never present unless the given a–v pair is present. Related pairs of a given a–v pair are those a–v pairs present some of the time when a given a–v pair is present (304). The system calculates relationships between a–v pairs to produce tables of a–v pairs presented according to the relationships (304, 310). The user performs additional analysis by investigating the a–v pair relationships through a graphical users interface. Additional visualization of the data are possible such as through Venn diagrams and animations. Plain–text data documents collected, for example from the Internet can be analyzed. In this case, the system pre-processes the text data to build a–v pairs based on sentence syntax.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

DATABASE ANALYSIS USING A PROBABILISTIC ONTOLOGY

BACKGROUND OF THE INVENTION

This invention relates in general to computer database systems and more specifically to a computer database system using an ontology structure to allow analysis of the database.

The proliferation of computer systems and improvements in telecommunications makes an overwhelming amount of data available to a computer user. Massive networks such as the Internet provide millions upon millions of data items in the form of words, numbers, images, etc., in very diverse and unregulated formats. Other, smaller, databases and database systems, such intranets and stand-alone computer systems, are more restrictive in their data formats yet still provide large volumes of data to the user. Perhaps, the smallest application of a computerized database is with today's so-called personal digital assistance (PDAs) which may contain an individual's address book, calendar, or similar personal database.

Within the range of all of these database systems lie the same basic problems of efficient access to, and analysis of, the data. Typical database applications are designed primarily to provide ease of data entry, upkeep and retrieval. However, the applications require that the database be specifically designed for a target application, e.g., medical record-keeping, so that "records," "templates," or similar structures must be designed by a database programmer or architect in order for the database application to be useful to an end user. For example, the Access database program, manufactured by Microsoft Inc., requires the creation of records having multiple fields. Information of predetermined types is entered into the fields. The information is accessed using the same predetermined fields.

Such a database system provides a query language so that a user can form relational inquires into the database. In a medical records database, for example, a user can retrieve records from the database that include a specific patient's name AND were created AFTER a certain date. The "AND" operator is a relational operator between the

two desired attributes of "patient name" and "creation date." The AFTER operator modifies the query range by using the creation date.

More recently, popular database search engines have been created which allow users to search larger, less-structured databases such as the Internet with similar relational query operators. For example, search engines by Yahoo! and AltaVista allow relational queries using keywords that do not relate to specific attributes and fields. Instead, any document having words with a specified relationship, such as the above relationship used in the example, is listed as a possible document of interest to the user. While relational database queries are useful in searching for information, they require that the user know with high specificity the type of information sought.

Another use for databases is to provide a platform for analyzing data to determine characteristics, trends or predictive guidelines in the information. For example, where financial data is being analyzed it may be useful to discover that where inflation is high in an overseas market, bond prices in a different market are also high with a very high frequency. Or, in a medical research application, it would be useful to determine that in a high percentage of cases where a certain treatment was used the recovery time was very short. However, such analysis of data is not possible with traditional database applications which singularly focus on retrieving existing data and entering and maintaining data of predetermined formats using relational queries.

Thus, it is desirable to provide a technique and system for analyzing characteristics of data in the matter discussed above. Further, it is desirable to provide such a technique and system that is usable with databases regardless of the size or level of structuring of the database. Also, given the vast amount of data available, it is vital that the results of the analysis be presented in a form that is efficient for detecting trends, qualities or other useful information among the data being analyzed.

SUMMARY OF THE INVENTION

The present invention provides a method and system for efficiently analyzing databases. In one embodiment, the invention is used to analyze data represented in the form of attribute-value (a-v) pairs. A primary step in building the ontology is to identify parent, child and related a-v pairs of each given a-v pair in the database. A parent is an a-v pair that is always present whenever a given a-v pair is

present. A child is an a-v pair that is never present unless the given a-v pair is present. Related pairs of a given a-v pair are those a-v pairs present some of the time when a given a-v pair is present.

The system calculates relationships between a-v pairs to produce tables of a-v pairs presented according to the relationships. The user performs additional analysis by investigating the a-v pair relationships through a graphical user interface. Additional visualizations of the data are possible such as through Venn diagrams and animations. Plain-text data documents collected, for example, from the Internet can be analyzed. In this case, the system pre-processes the text data to build a-v pairs based on sentence syntax.

One embodiment of the invention provides a method for analyzing a database where the database includes a plurality of records having a-v pairs. The method executes on a computer system and includes the following steps: determining two or more parents of a given a-v pair where a parent of an a-v pair is another a-v pair that exists within every record that the given a-v pair exists; and displaying the two or more parents on the display screen along with an indication that two or more parents are associated with a given a-v pair.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 shows a basic computer system suitable for executing the software of the present invention;

Fig. 2 shows subsystems in the computer system of Fig. 1;

Fig. 3 shows a generalized computer network;

Fig. 4A shows a taxonomy diagram;

Fig. 4B shows an ontology diagram;

Fig. 5 is an example of a Table-Display of the present invention;

Fig. 6 is a first Venn Display;

Fig. 7 is a second Venn Display;

Fig. 8A is a first frame of a Venn Display animation;

Fig. 8B is a second frame of a Venn Display animation; and

Fig. 8C is a third frame of a Venn Display animation.

DESCRIPTION OF THE SPECIFIC EMBODIMENTS

Fig. 1 shows a basic computer system 1 suitable for executing the software of the present invention. Computer system 1 includes a display device 3 with a display screen 5. Cabinet 7 houses additional components of the computer system (not shown) such as a processor, memory, disk drive, compact disc read only memory (CD-ROM), etc. Keyboard 9 and mouse 11 are standard user input devices. Mouse 11 includes buttons 13 for facilitating user input.

Fig. 2 shows several subsystems interconnected via a bus 115. Central processor 108 communicates with system memory 107, fixed disk 104, I/O controller 106, display adapter 111, serial port 102 and netowork interface 105. some subsystems interface to user input and output devices such as the monitor, remote pointing device (RPD) or, "mouse," and keyboard which are also shown in Fig. 1. Network interface 105 is used to connect the computer system to additional external computer systems on a network. Many other configurations of subsystems are possible. A computer system suitable for use with the present invention can use less subsystems, components or devices than those shown in Fig. 2. For example, a handheld computer may include only a processor, memory (both random-access memory (RAM) and read-only memory (ROM)), small display screen and keyboard. Also, computer systems suitable for use with the present invention may include more subsystems than are shown in Fig. 2. For example, the handheld computer may include a PCMCIA card interface for storing and retrieving data from an external card.

Fig. 3 shows a generalized computer network.

In Fig. 3, servers such as server 150, 151 and 152 may be interconnected by any communication means and in any network topology. For example, the servers can be connected by hardwire, radio frequency transmissions, infrared transmissions, etc. They can be connected in a star, ring, daisy chain, etc., schemes. Also, any communication protocol such as Ethernet, IEEE 1394 or TCP/IP can be employed.

User stations 153, 154 and 155 are shown connected to server 151. Again, the interconnection between the computer stations and the server, or servers, can be by any means, topology and protocol as is known. Although all of the computer stations are shown connected to server 151, additional servers having any number of additional computer stations can be added to the interconnected network of Fig. 3. Although the

specific embodiment is discussed with reference to a single computer station, such as computer station 153, accessing a database on a server, such as server 151, it should be readily apparent that the database can be distributed over two or more servers or computers. Further, the database can exist partially, or completely, on the computer stations themselves. That is, computer station 153 can operate as a stand-alone database where the database engine, along with the data, auxiliary programs, etc. all reside within the computer station. The idea of distributed databases is well-known in the art. Many variations on distributing data, and many designs of database "front ends" or user interfaces are possible without deviating from the scope of the present invention.

A preferred embodiment of the invention executes software on a workstation. The workstation can typically contain the database that is being analyzed. Alternatively, the workstation can be connected to a network and the database can be remote. A user operates the software via the user input devices. Output is presented to the user through the display screen or by other methods such as hard copy output from a printer. It should be apparent that, although not directly discussed here, many types of user input and output devices are suitable for use with the present invention. For example, a trackball, digitizing tablet, light pen, data glove,etc. can all be used to provide input to the software.

Fig. 4A shows a taxonomy diagram for a classification system used in zoology. Fig. 4A is presented and discussed to introduce the concept of an ontology (a generalized taxonomy) and to define some terms.

In Fig. 4A, the top level, or row, of the diagram is the "familiy" while the next two rows down are the "genus" and "species." This information can be represented in a database as "entries" consisting of attribute-value pairs. This database representation is shown in Table I, below.

Family		Genus	Species
1	Micrococcaceae	Staphylococcus	Aureus
2	Micrococcaceae	Staphylococcus	Saprophyticus
3	Micrococcaceae	Staphylococcus	Epidermidis
4.	Micrococcaceae	Micrococcus	Luteus

Table I

In Table I, the columns labelled "Family," "Genus" and "Species" are attributes. Each horizontal row is an entry in the database. Each entry has values for each defined attribute as shown in the corresponding column. Thus, entry 1 has a-v pairs as follows: (Family, Micrococcaceae), (Genus, Staphylococcus) and (Species, Aureus). These can be abbreviated as (F, M), (G, S) and (S, A).

A feature of the present invention is the ability to "work backward" to determine classification schemes based on database entries such as those shown in Table I. The system defines "parent" and "child" a-v pairs in relation to other

a-v pairs. A first a-v pair is a parent of a second a-v pair if the first a-v pair occurs in every entry that the second a-v pair occurs. Thus, (F, M) is a parent to (G, S), (G, M), (S, A), (S, S), (S, E) and (S, L). A first a-v pair is a child to a second a-v pair if the first a-v pair never occurs in an entry unless the second a-v pair is also in the entry. From Table I, (G, S) is a child of (F, M) and (S, A) is a child to both (G, S) and (F, M).

By starting with a database, such as the database represented in Fig. 4A, an analysis can be performed on the a-v pairs to determine all of the parent and child relationships. By considering parent pairs as classes of those pairs that are their childs, the classification hierarchy shown in Fig. 4A is achieved.

A characteristic of the data shown in Fig. 4A and Table I is that each a-v pair has, at most, one parent a-v pair. It is easy to imagine databases where more than one parent exists for a given a-v pair. Such a database example is shown in Fig. 4B where the item CAR has more than one parent. Note that, in Fig. 4B, the items will be treated as values of attributes. The attributes themselves are not named but can be assigned as "quality" for the top row, "object" for the second row and "manufacturer" for the bottom row. In this case, the a-v pair (object, car) has (quality, transport vehicle) and (quality, collector's items) as its two parent nodes. Such an organization of data where an item can have more than one parent is referred to as an "ontology."

A generalization to the ontology organization is to allow probabilistic relationships between the a-v pairs. So far, parent and child a-v pairs are shown as absolute existences. However, in any database, especially large databases, there are likely to be errors in the data. Also, characteristics and trends of interest will likely show up as statistical occurences of something less than 100%. The ontology described so far is not

flexible in handling rates of occurence. The present invention solves this problem by creating a probabilistic ontology where statistics on rates of occurence of parent and child relationships are computed and compiled for use in analysis.

A preferred embodiment of the present invention is referred to as the "High-Performance Ontology Builder and Browser" (HOBB). HOBB not only generates an ontology but it also allows the user to "browse" attribute-value pairs that intersect in terms of common occurences in database entries, but that aren't in strict parent/child relationships to each other. In other words, their parent/child relationships need not occur at 100%.

Fig. 5 is an example of the Table Display of HOBB. The values displayed are part of an analysis of a film database from the Human-Computer Intereaction Group at UMD. The database provided about 1750 entries, or records, of 9 attribute-value (a-v) pairs each. Each database entry includes an a-v pair of a film title, subject, length, actor, actress, director, popularity, awards, and year of make. In figure 5, HOBB is analyzing film data and is presenting the results of calculating parent/child relationships to the user.

Center column 304 includes, at the top, the a-v pair of interest at 306. This is listed as "Actor, Martin Steve." To the right of the a-v pair of interest, at 308, is the number of times that the pair occurred in the database, which in this case is nine. Left column 302 lists parents of "Actor,Martin Steve". As expected, "Subject,Comedy" and "Awards,No" have been detected as parents. Even though the matrix of data is fairly populated in this example, the same sort of capability would have been detected in even a sparse matrix of data. Right column 310 lists children of "Actor,Martin Steve". In other words, every time "Director,Reiner Carl" appeared I the database, "Actor,Martin Steve" also was there.

The entries in center column 304 below "Actor,Martin Steve", the a-v pair of interest, are a-v pairs that co-occurred in the database with "Actor,Martin Steve", ranked in order of highest frequency. The frequency of occurrence as a percentage is listed to the left of each pair. For example, one item in center column 304 is "Year,1987". This indicates that the year 1987 appeared with 22.2% (i.e. 2 of the 9) of the films in the database where "Actor,Martin Steve" appeared. We also see that "Length,60" co-occurred with "Actor,Martin Steve" 22.2% of the time. These related pairs are neither parents nor children of the a-v pair of interest, but may provide insight into the data because of their rather large "overlap" of occurrence with the a-v pair of

interest. In this case, the two movies of curious titles "Steve Martin The Funnier Side of Eastern Canada" and "Steve Martin Live" were both exactly 60 minutes, and thus were probably TV specials. By displaying this portion of the probabilistic ontology the system of the present invention allows a user to quickly make inferences and form theories about relationships between the data.

Refinements to the user interface are possible. For example, the system can allow the user to specify a cut-off threshold below which related pairs will not be displayed. In Fig. 5, where the cut-off to be set to "above 15%", those pairs below "Director,Reiner Carl" would not be displayed. Also, thresholds can be applied to parent and child criteria so that 100% co-occurrences are not required to place a pair into the parent or child column for a given a-v pair.

In the preferred embodiment, all of the a-v parent, child and co-occurrence relationships are pre-computed. This allows instant display of user interrogations into the a-v relationships. For example, a user can mouse-click on "Director,Reiner Carl" either in the middle or the right column to make "Director, Reiner Carl" the a-v pair of interest. "Director,Reiner Carl" will then be displayed at 306 and the display will update to show all related a-v pairs to "Director,Reiner Carl". This "browsing" feature of HOBB is very useful to the researcher in discovering relationships. The browsing feature is all the more useful because the display updating, when a new a-v pair of interest is selected, is instantaneous due to pre-computing. This allows a user to maintain concentration, be more efficient, and investigate a large number of possible relationships.

Another advantage of computing the a-v relationships is that there is no need to keep the original database with the relationships database. The relationships database may be much smaller than the original database. For example, where only a few attributes from each entry are of interest the entire entry need not be analyzed and the resulting relationships database can be smaller than the original database. Also, there may be security issues in copying the original database in its original form. Once the relationships database is created it can be analyzed separately from any hardware and software necessary to support the original database.

Yet another implementation of the invention uses existing database programs to examine the ontology. Once an ontology database of a-v pairs and their parent, child, co-occurrence relationships is created, the ontology database can simply be fed as data to an off-the-shelf database application program such as Lotus Excel or

Microsoft Access. The user can operate these databases using the traditional controls provided by the third party database manufacturer, or the user can design a customized front-end to approximate the functions of the HOBB program presented herein. This allows the system of the present invention to be adaptable to small computers, such as personal computers, with a minimum of effort.

An example of applications where HOBB can assist a database researcher is where an economist has a database where each quarter is an entry, and within these entries are a-v pairs to keep track of Gross Domestic Product (GDP) growth, exports, market movements, bank lending, etc., with all applicable leads and lags in time. Using the system of the present invention (Exports, High) can be selected as the a-v pair of interest. This might show that every time (GDP Growth, High), then (Exports, High) occurs two quarters later. In other words, (GDP Growth, High) is a parent to (Exports, High). Also, the same screen might showthat every time (Exports, High), then (Consumer Confidence, Low), meaning that (Exports, High) is a parent to (Consumer Confidence, Low). By browsing around, relationships between economic occurrences will begin to form and the ones that seem prominent can be researched theoretically and otherwise, resulting in a much better understanding of the economy from a simple database.

Some additional examples of HOBB's utility could be seen in the following professions of Table II:

1.***Economist*** – After a bit of work, you were able to gather time series dating back to 1950 on a quarterly basis covering France's and Italy's market and economic movements. Given this spreadsheet of 188 rows and numerous columns, what sort of information would be most valuable? You would be interested in questions like: When the French GDP is shrinking, what tends to happen in the Italian series? Do Italian interest rates seem affected? What about exports? All these questions are co-occurrence questions which HOBB, through its browsing feature, makes clear and explorable.

2.***Medical Researcher*** – Over the period of a year, your hospital has been keeping track of infections, how they were treated, and how successful the treatment was. The result is a large file of patient names along with bacteria names; antibiotic names, dosages, and days of use; and perceived side effects. Do certain side effects coincide

with certain antibiotic dosages? If Vancomycin is ineffective, what other antibiotics tend to be ineffective? Are there certain antibiotics that do not work well with Genus Pseudomonas? All these questions are co-occurrence questions which HOBB, through its browsing feature, makes clear and explorable.

3.*Retail Marketer* – After much trouble and expense, your grocery chain has set up a tracking system that records each grocery purchase and stores the information in an Oracle database. You are now setting up a new store in a busy area of town and you want to convert your newfound data on grocery purchases into a layout that maximizes convenience for your customers. When customers purchase mayonnaise, how often is this accompanied by pickles? If they get peanut butter, do they also always buy bread? HOBB is a way to browse the data and get a solid understanding of grocery purchases before you take pen in hand and lay out the shelves.

4.*Direct Marketer* – Using a combination of databases, you gather demographic data on 50,000 customers you feel are good candidates for your mailings. After sending a test mailing, you would like to see if there is some consistent elements or combination of elements between the demographic data and whether or not the customers responded. Once again we need to see the database through the lens of co-occurrence, which can be done utilizing HOBB.

TABLE II

The system of the present invention can be adapted for use in more generalized databases that are not already represented as a-v pairs. In these cases, the database is first pre-processed to generate the a-v pairs. For example, in a text database, such as documents from the Internet, each document is treated as a record or entry. The occurrence of a given word in a sentence, as well as co-occurrences of other words with the given word in each sentence, is used to build the a-v pairs which are analyzed by the system.

A feature of the present invention is the ability it provides to "visualize" data. Although the table display of Fig. 5 provides an adequate interface for looking at precise relationships between data, it requires some work and scrutiny do determine more "global" relationships involving larger number of a-v pairs. For example, where a first a-v pair co-occurs at 80% with another a-v pair it would seem to imply that there is a strong

relationship with the two pairs. However, If the first a-v pair also occurs in 98% of the entries in the database then the fact that it intersects at 80% with the other a-v pair is not as significant. In fact, it becomes significant that it intersects with the other a-v pair only 80% of the time! In order to determine this from the table display of Fig. 5, a user must not just detect the, seemingly, high co-occurrence of the pairs in the middle column, but must compare the occurences of each a-v pair to the database as a whole.

To provide better global analysis of relationships the invention uses Venn diagrams in a "Venn Display" to show co-occurences of a-v pairs as overlaps in the diagrams. By presenting co-occurrences visually it is easier to detect strong relationships between data.

Figs. 6 and 7 show two examples of Venn Displays. These diagrams are displayed in color in the actual system. Fig. 6 shows (Inflation, High) as the attribute value pair in yellow, and represented by yellow circle 350. This is the a-v pair selected, or designated as "of interest," such as the pair displayed at the top of the center column in Fig. 5, as discussed above. The pair (Long Bond Rates, High) is a second pair designated by the user for comparison with the pair of interest. In the preferred embodiment, the user can mouse-click on any a-v pair on the table display of Fig. 5 to designate the clicked pair for comparison. The user can use the scroll bars to the right of each column to bring additional pairs into view.

Fig. 6 shows (LBR, H) as blue circle 352. The relative sizes of each circle, along with their area of overlap 354, are proportional with respect to the number of occurences. That is, (I, H) occurs 20 times in the database and has a yellow circle 350 that is about 2/3 of the area of the blue circle 352 representing (LBR, H) which occurs 30 times in the database. The area of overlap of the two circles is 18, which is the number of times that the two a-v pairs co-occur in the database entries. Using the Venn Display of Fig. 6, the user can quickly see co-occurrence relationships and is prevented from making errors of the type discussed above where a true interpretation of data relationships hinges on an idea of the percentage of occurence of each a-v pair to the entire database.

Fig. 7 shows a second form of Venn Display, the "Full Information Display." Using the prior example of an economic model, assume that the attribute "Inflation" can have one of three values, either "Low," "Medium" or "High." The interaction between (LBR, H) and Inflation for every possible inflation value is shown graphically in Fig. 7. It is easy to see that most of the occurences of (LBR, H) are when

(I, H). Also, the overlap of (LBR, H) with (I, H) is a larger percentage of the overall occurences of (I, H) in the database than with the other a-v pairs. That is, (LBR, H) occurs in 5/14 occurrences of (I, H); (LBR, H) occurs in 2/10 occurrences of (I, M) and (LBR, H) occurs in 1/16 occurrences of (I, L). Again, while not shown in Fig. 7, color is used to designate each of the regions (I, L), (I, M) and (I, H).

Figs. 8A-C show frames of a "movie" formed of several Venn Displays to create an animation that illustrates a change in data over time.

Suppose a researcher is interested in using the database to see if Streptococcus Fataliti is developing resistance to the antibiotic Vancomycin. A display similar to that of Fig. 7, the "Full-Information Display" is computed over different time intervals. These are shown in succession at a desired speed. From the movement of the center circle over time, it can be seen that the bacteria are gaining resistance to Vancomycin (i.e., it takes longer periods of treatment with Vancomycin to kill the bacteria).

Thus, a system for analysis and visualization of data has been presented. Although the invention has been discussed with respect to a specific embodiment, many modifications to the specific embodiment are possible without deviating from the invention, the scope of which is determined solely by the appended claims.

WHAT IS CLAIMED IS:

1. A method for analyzing a database, the database including a plurality of records having attribute-value (a-v) pairs, the method executing on a computer system having a processor and display screen, the method comprising the following steps:

determining, by using the processor, two or more parents of a given a-v pair, wherein a parent of an a-v pair is another a-v pair that exists within every record that the given a-v pair exists; and

displaying the two or more parents on the display screen along with an indication that the two or more parents are associated with the given a-v pair.

2. A method for analyzing a database, the database including a plurality of records having attribute-value (a-v) pairs, the method executing on a computer system having a processor and display screen, the method comprising the following steps:

determining, by using the processor, a probable parent of a given a-v pair, wherein a probable parent of an a-v pair is another a-v pair that exists within a subset of the records that the given a-v pair exists; and

displaying the probable parent on the display screen along with an indication that the probable parent is associated with the given a-v pair.

3. The method of claim 2, further comprising the step of

using a threshold value to determine the frequency of co-occurence of a candidate parent with an a-v pair that is necessary before the candidate parent is displayed as a probable parent.

4. The method of claim 3, the computer system coupled to a user input device, the method further comprising the step of

accepting signals from the user input device to specify the threshold value.

5. The method of claim 2, further comprising the step of

precomputing parent and child relationships.

6. The method of claim 2, the computer system coupled to a user input device, the method further comprising the steps of

displaying a plurality of a-v [pair] pairs on the screen;

accepting signals from the user input device to select an a-v pair; and

displaying parent and child relationships with respect to the selected a-v pair.

7. An apparatus for analyzing a database, the apparatus comprising:

a computer system including a display coupled to a processor;

a database coupled to the computer system, the database including a plurality of records having attribute-value (a-v) pairs;

one or more computer instructions for determining two or more pairs of a given a-v pair, wherein a parent of an a-v pair is another a-v pair that exists within every record that the given a-v pair exists; and

one or more computer instructions for displaying the two or more parents on the display along with an indication that the two or more parents are associated with the given a-v pair.

8. One or more computer instructions stored in a computer-readable medium for performing the steps of claim 1.

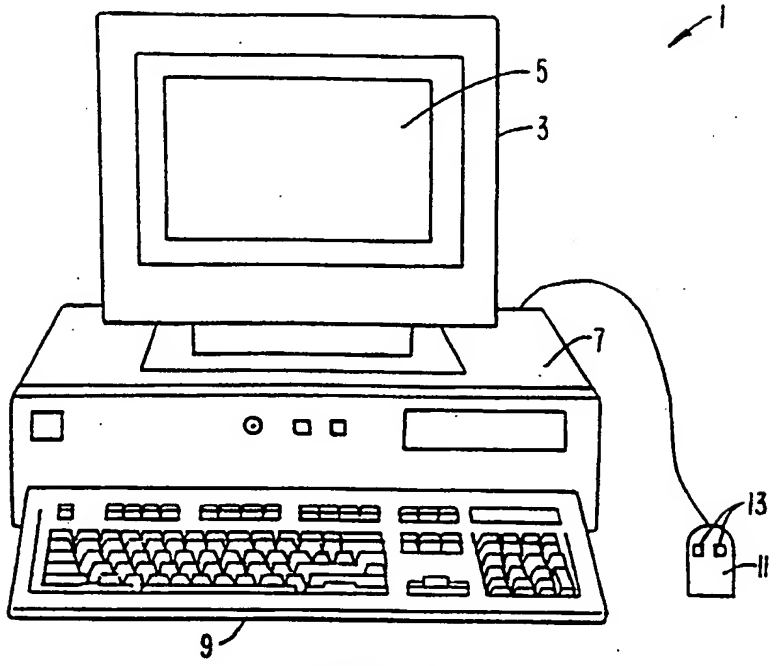

1
5
3
7
13
11
9

FIG. 1.

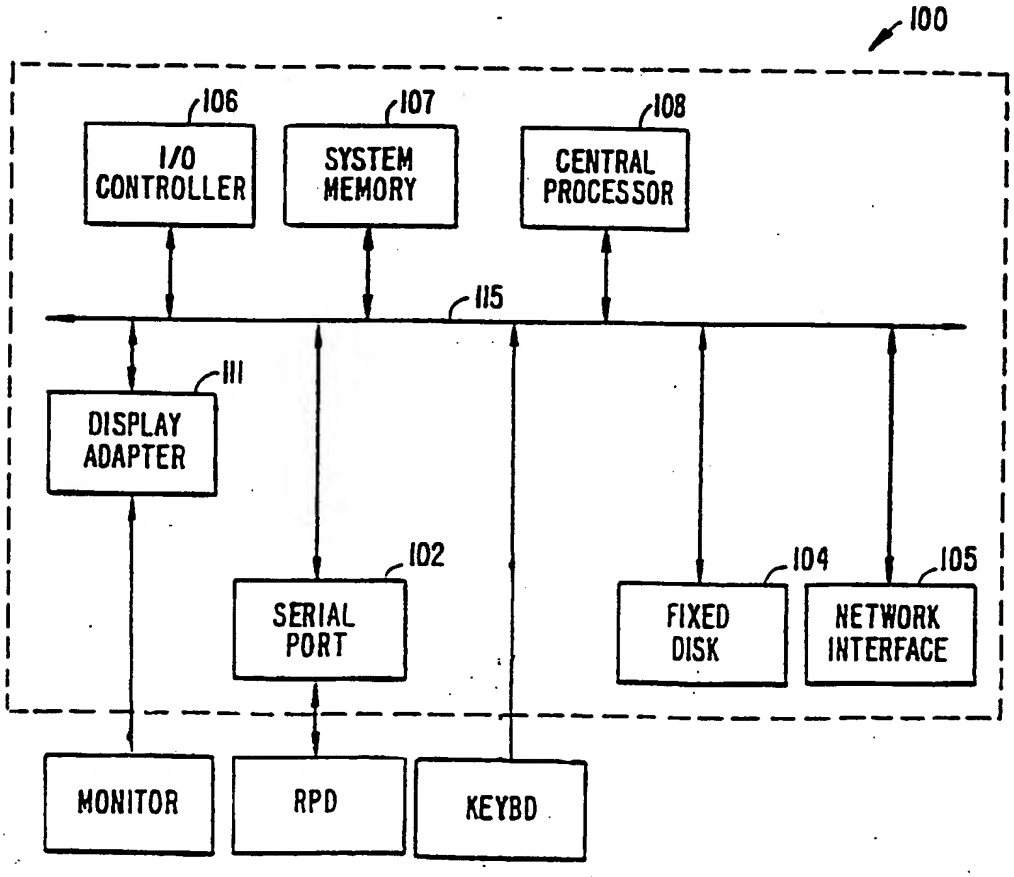

100
106
I/O CONTROLLER
107
SYSTEM MEMORY
108
CENTRAL PROCESSOR
115
111
DISPLAY ADAPTER
102
SERIAL PORT
104
FIXED DISK
105
NETWORK INTERFACE
MONITOR
RPD
KEYBD

Fig. 2

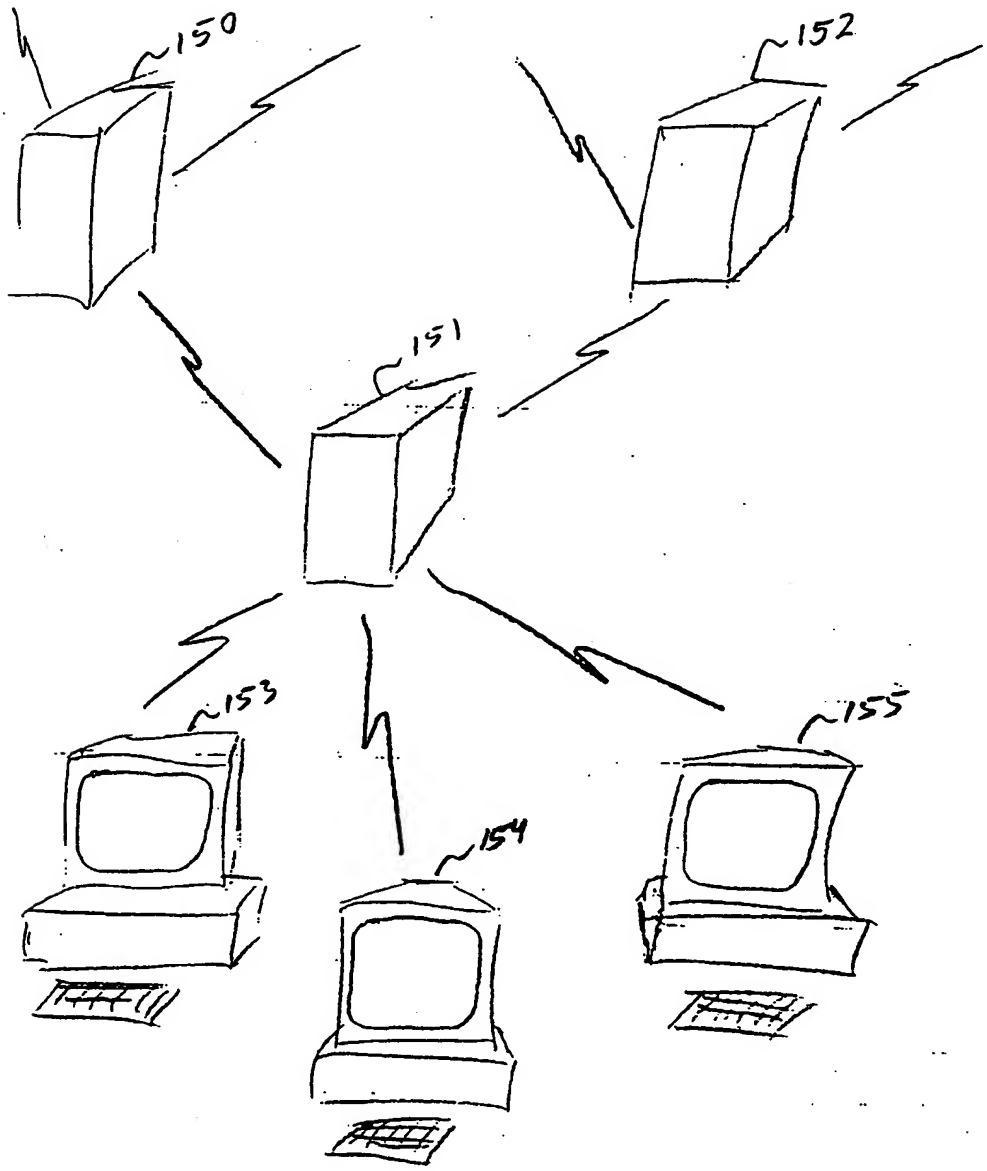


Fig. 3
2/9

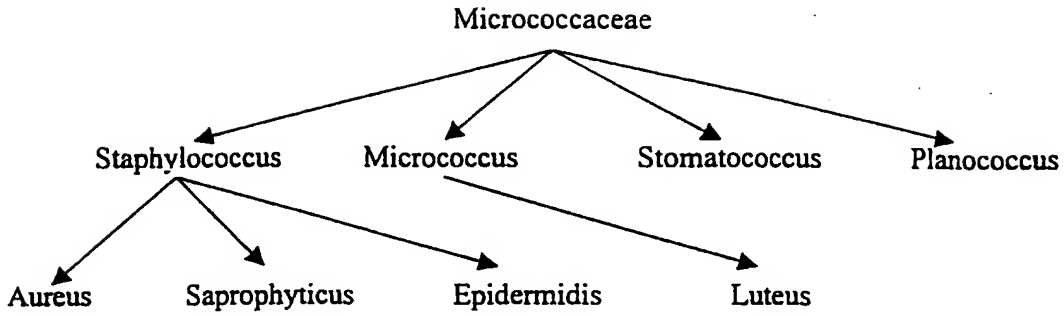

Micrococcaceae
Staphylococcus
Micrococcus
Stomatococcus
Planococcus
Aureus
Saprophyticus
Epidermidis
Luteus

Fig. 4A

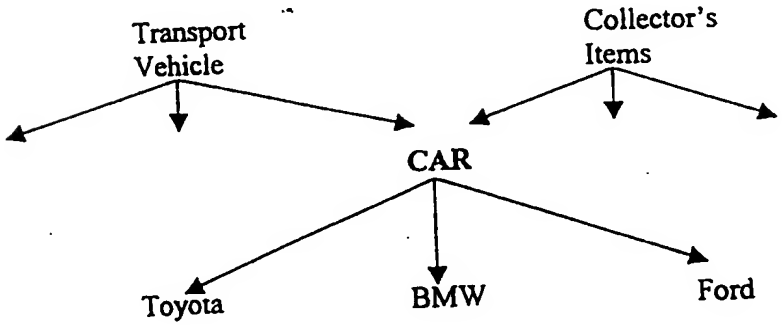

Transport Vehicle
Collector's Items
CAR
Toyota
BMW
Ford

Fig. 4B

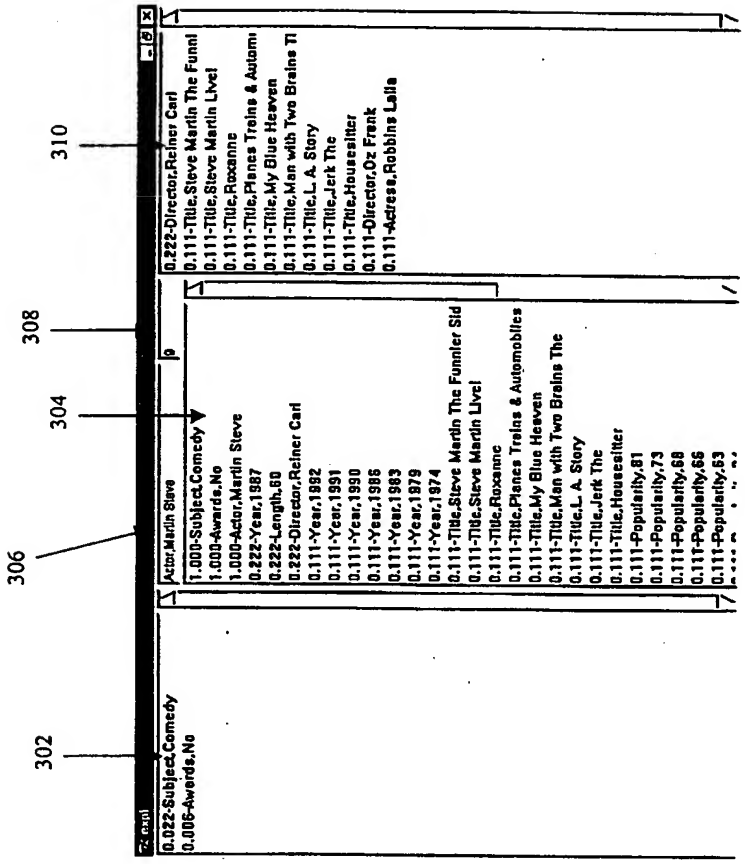

302
306
304
308
310
0.022-Subject,Comedy
0.006-Awards,No
Actor,Martin Steve
1.000-Subject,Comedy
1.000-Awards,No
1.000-Actor,Martin Steve
0.222-Year,1987
0.222-Length,80
0.222-Director,Reiner Carl
0.111-Year,1982
0.111-Year,1991
0.111-Year,1990
0.111-Year,1986
0.111-Year,1983
0.111-Year,1979
0.111-Year,1974
0.111-Title,Steve Martin The Funnier Sid
0.111-Title,Steve Martin Live!
0.111-Title,Roxanne
0.111-Title,Planes Trains & Automobiles
0.111-Title,My Blue Heaven
0.111-Title,Man with Two Brains The
0.111-Title,L. A. Story
0.111-Title,Jerk The
0.111-Title,Housesitter
0.111-Popularity,81
0.111-Popularity,73
0.111-Popularity,68
0.111-Popularity,66
0.111-Popularity,63
0.222-Director,Reiner Carl
0.111-Title,Steve Martin The Funni
0.111-Title,Steve Martin Live!
0.111-Title,Roxanne
0.111-Title,Planes Trains & Automi
0.111-Title,My Blue Heaven
0.111-Title,Man with Two Brains Tl
0.111-Title,L. A. Story
0.111-Title,Jerk The
0.111-Title,Housesitter
0.111-Director,Oz Frank
0.111-Actress,Robbins Laila

Fig. 5

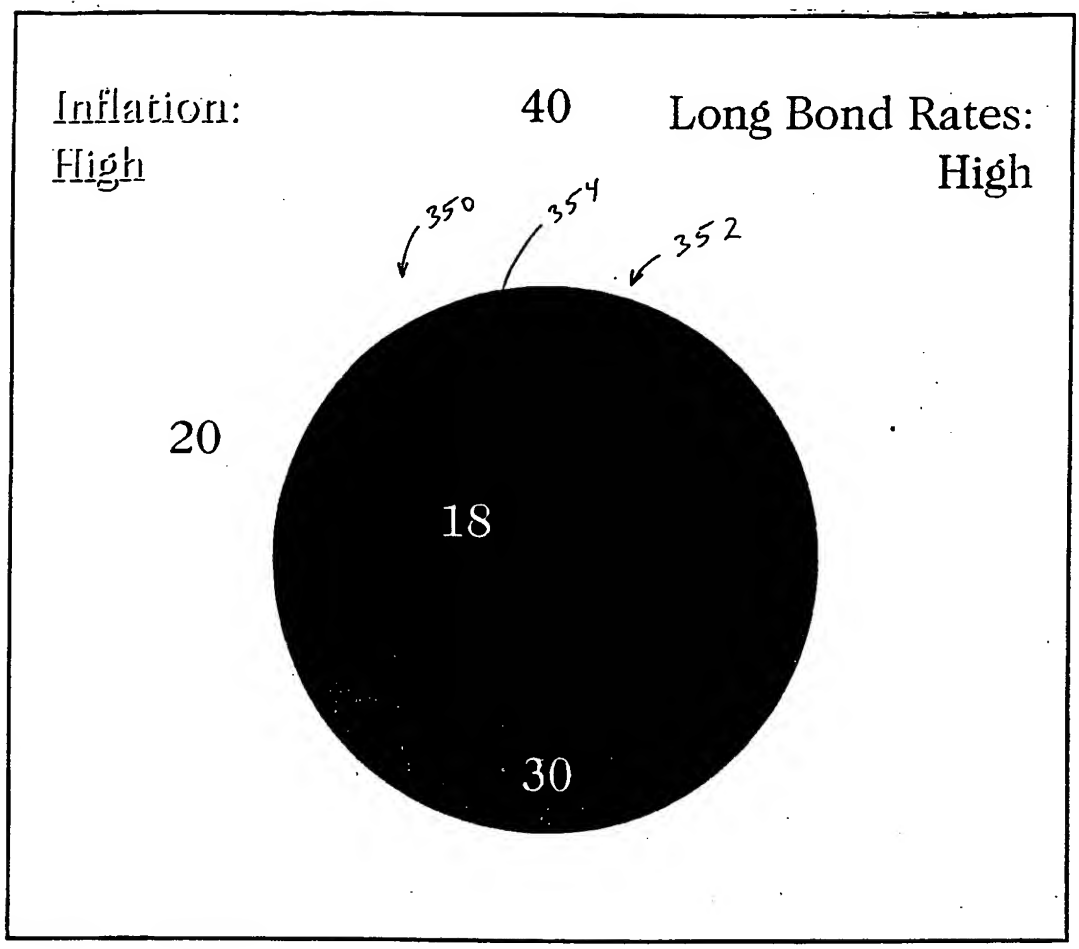


Fig. 6

Venn Display

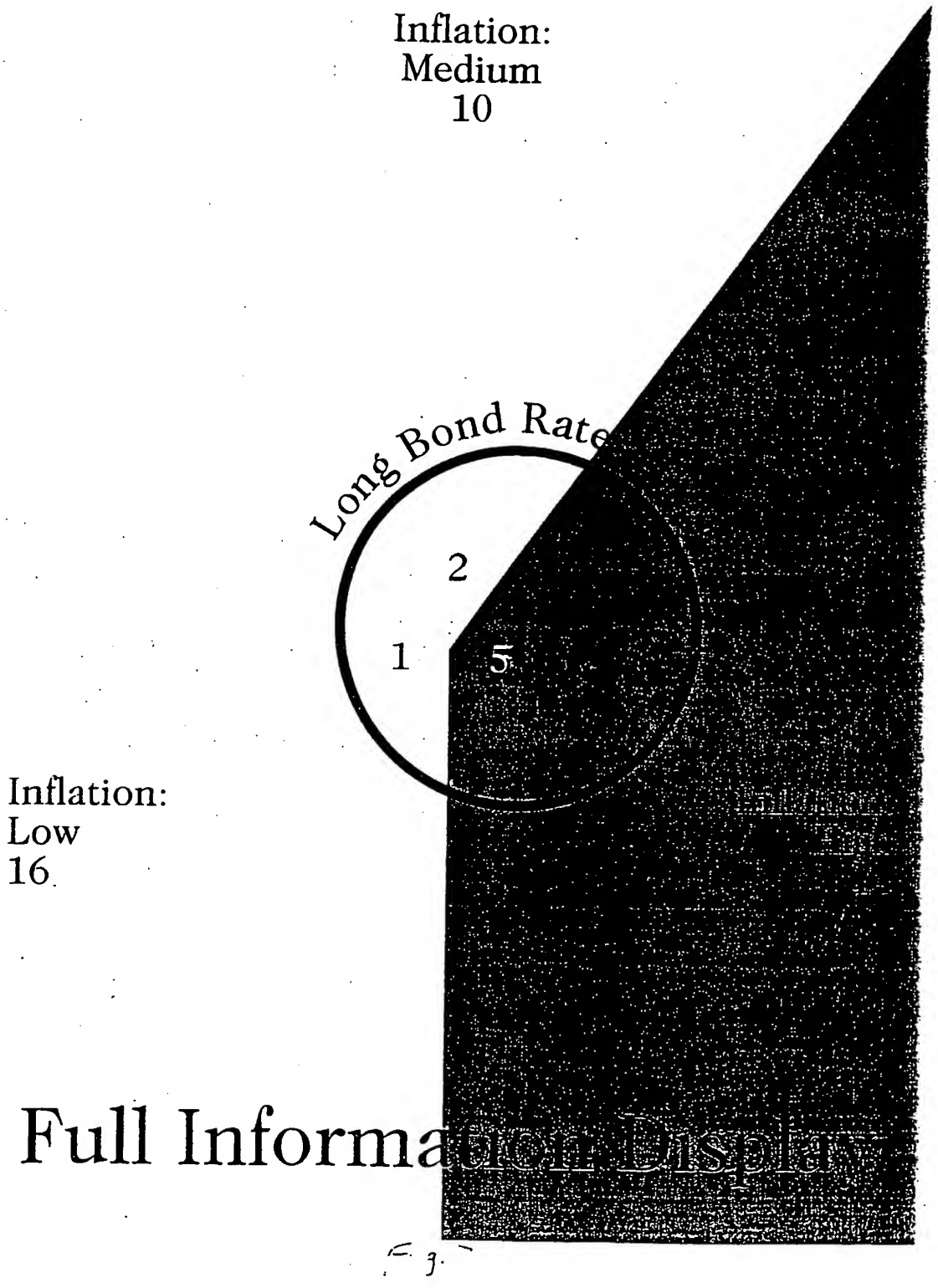

Inflation:
Medium
10
Long Bond Rate
2
1
5
Inflation:
Low
16
Full Information Display

Fig. 3

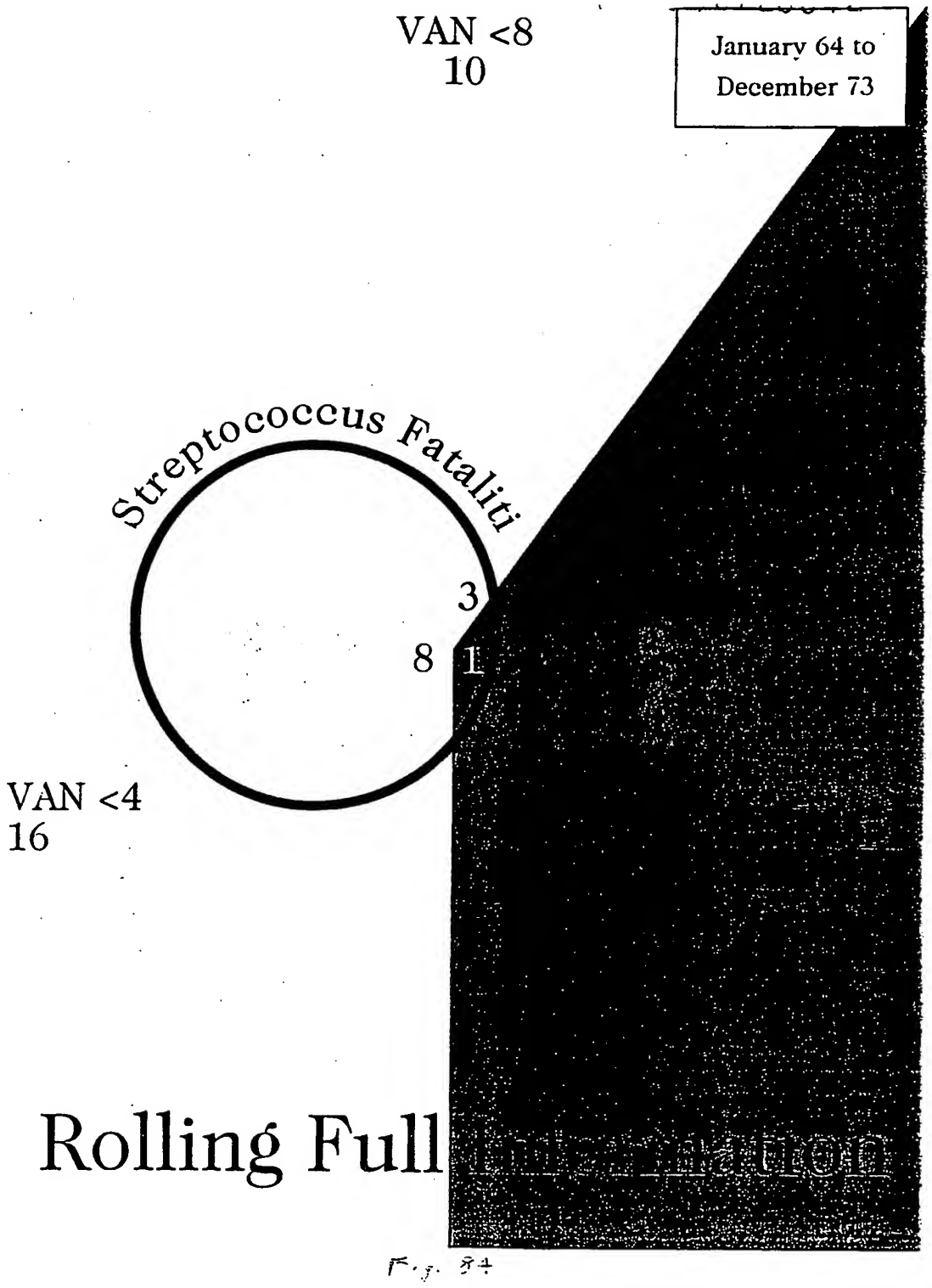
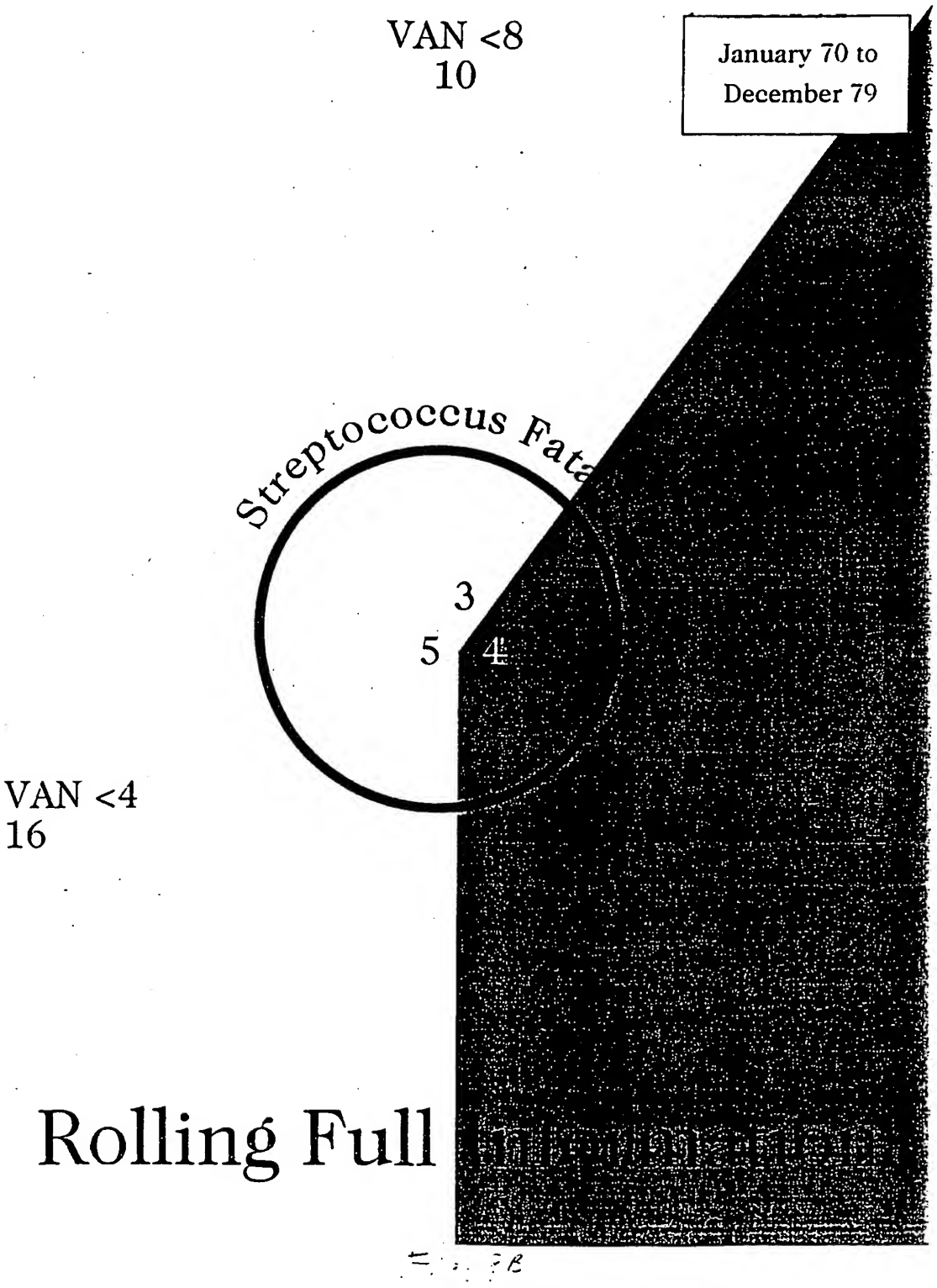

VAN <8
10
January 64 to
December 73
Streptococcus Fataliti
3
8
1
VAN <4
16
Rolling Full

Fig. 3A


VAN <8
10
January 70 to
December 79
Streptococcus Fata
3
5
4
VAN <4
16
Rolling Full

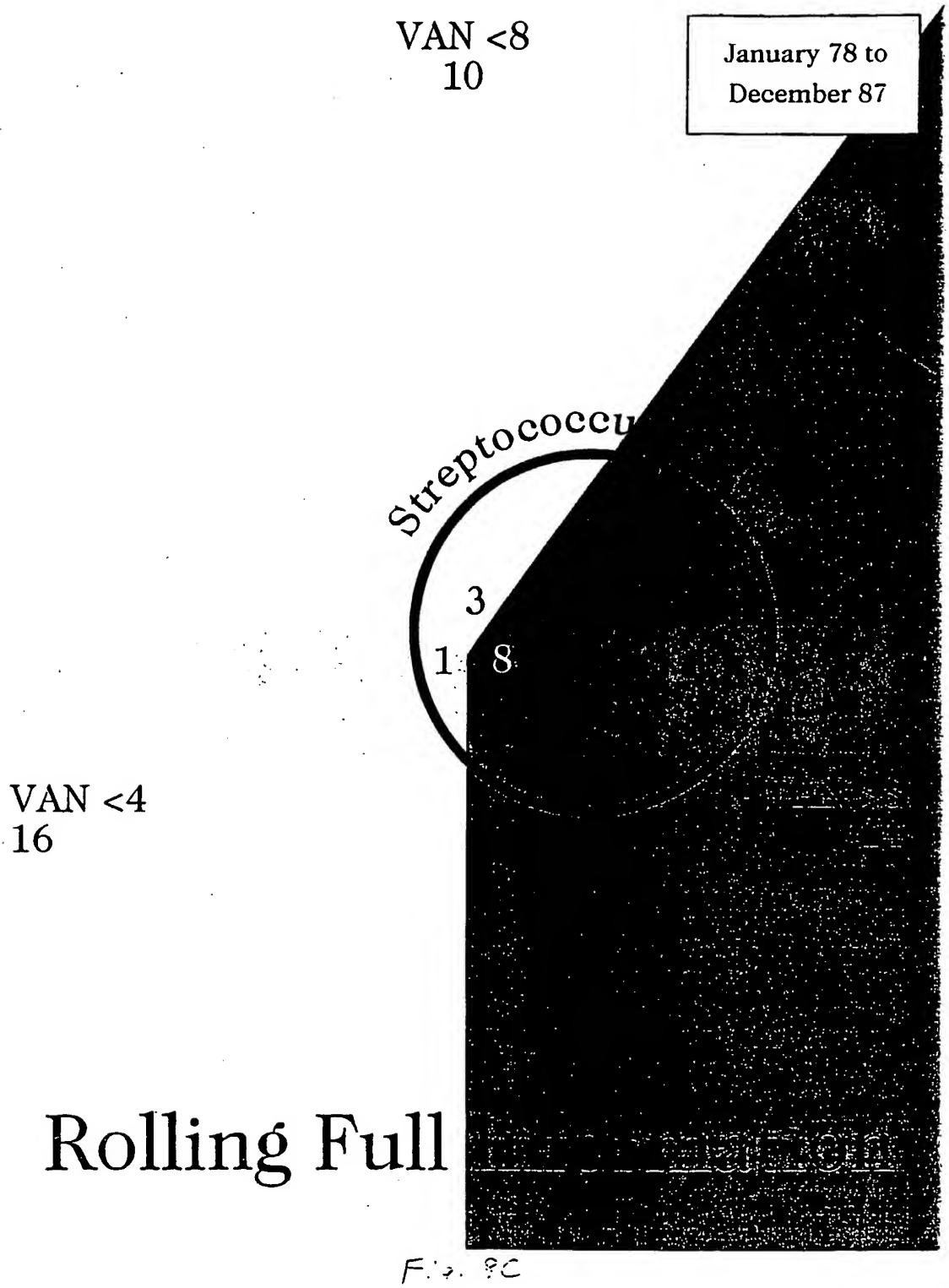

VAN <8
10
January 78 to
December 87
Streptococcu
3
1
8
VAN <4
16
Rolling Full

Fig. 9C

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US98/26642

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :G06F 17/30
US CL : 707/1

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 707/1

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

Please See Extra Sheet.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	RANADE, J. et al. DB2 Concepts, Programing, and Design,1991, McGraw-Hill, pp. 63-94, especialy pages 68-69.	1, 2, 5, 7 and 8
A, P	US 5,802,254 A (SATOU et al.) 01 September 1998, col. 2.	1-8
A, P	US 5,724,573 A (AGRAWAL et al.) 03 March 1998, col. 3, line 45 through col. 6, line 18.	1-8.
A	Han J. et al, Data Discovery of Quanitative Rules in Relational Databases, IEEE Transactions on Knowledge and Data Engineering, Vol. 5, No. 1, Febuary 1993 pp. 29-40.	1-8

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search 28 FEBRUARY 1999	Date of mailing of the international search report 31 MAR 1999
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer JACK M. CHOULES Telephone No. (703) 305-9840

Form PCT/ISA/210 (second sheet)(July 1992)*

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US98/26642

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	Ziarko, W. et al. Discovering Attribute Relationships, Dependencies and Rules by Using Rough Sets, Proceedings of the Twenty-Eight Hawaii Intenational Conference on System Sciences. January 1995, Vol. 3, PP. 293-299.	1-8

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US98/26642

B. FIELDS SEARCHED
Electronic data bases consulted (Name of data base and where practicable terms used):

APS, DIALOG, IEL
search terms: data, database, mining, knowledge, discovery, field, column, value, attribute, parameter, record, Venn, diagram, display, relation, probability, statistic, pair, ontology, analysis, correlate.